Case Studies 2022L

Local-dependence and accumulated-local dependence profiles

Apr 28, 2022

Reminder

- The partial dependence profile (PDP) show the marginal effect one or two features have on the predicted outcome of a ML model (Friedman, 2001).
- A PDP plot can show whether the relationship between the target and a feature is linear, monotonic or more complex.
- The PDP at a particular feature value represents the average prediction if we force all data points to assume that feature value.
- Therefore, PDP assumes that the features for which the PDP is computed a variable are not correlated with other feature.

Problems

- Non-additive models
- Correlated predictors
- Interactions between predictors

Local dependence profile

The PDP does not consider the correlations between predictors, because it is calculated based on marginal distribution of predictors. To capture the effect of all predictors, LD profile is proposed by Apley and Zhu (2020) based on conditional distribution of the predictors.

LD for model f() and predictor X^j is defined as follows:

$$g_{LD}^{f,j}(z) = E_{\underline{X}^{-j}|X^j=z}\{f(\underline{X})^{j|=z}\}$$

LD is the expected value of the model predictions over the conditional distribution of $\underline{X}^{-j}$ given $X^j = z$, over the joint distribution of all predictors other than X^j conditional on the value of the latter variable set to z. On the other words, it is the expected value of the CP profiles for X^j over the conditional distribution of $\underline{X}^{-j} | X^j = z$.

The marginal probability is the probability of a single event occurring, independent of other events. A conditional probability is the probability that an event occurs given that another specific event has already occurred.

Accumulated local effect profiles

- Accumulated local effect (ALE) describe how features influence the prediction of a ML model on average.

- ALE plot is a faster and unbiased alternative to PDPs.

- The ALE differs from PDP in two aspects:

    - ALE shows the differences in prediction instead of the averages. For example, if ALE for X2=1 is -2, it means that when X2=1, the prediction is lower than the average prediction by 2 unit of y.

    - ALE averages the difference in a small neighborhood of a data point. PDP averages over all the data points.

ALE

ALE profile for model f() and predictor X^j is defined as follows:

$$g^j_{ALE}(z) = \int_{z_0}^{z} [E_{\underline{X}^{-j}|X^j=v}\{q^j(\underline{X}^{j|=v})\}]dv + c$$

where z_0 is a value close to the lower bound of the effective support of the distribution of X^j and c is a constant, usually selected so that $E_{X^j}\{g^j_{ALE}(X^j)\} = 0$

$q^j(\underline{x}^{j|=v})$ describes the local effect (change) of the model due to X^j, or describes how much the CP profile for X^j changes at the intervals.

ALE

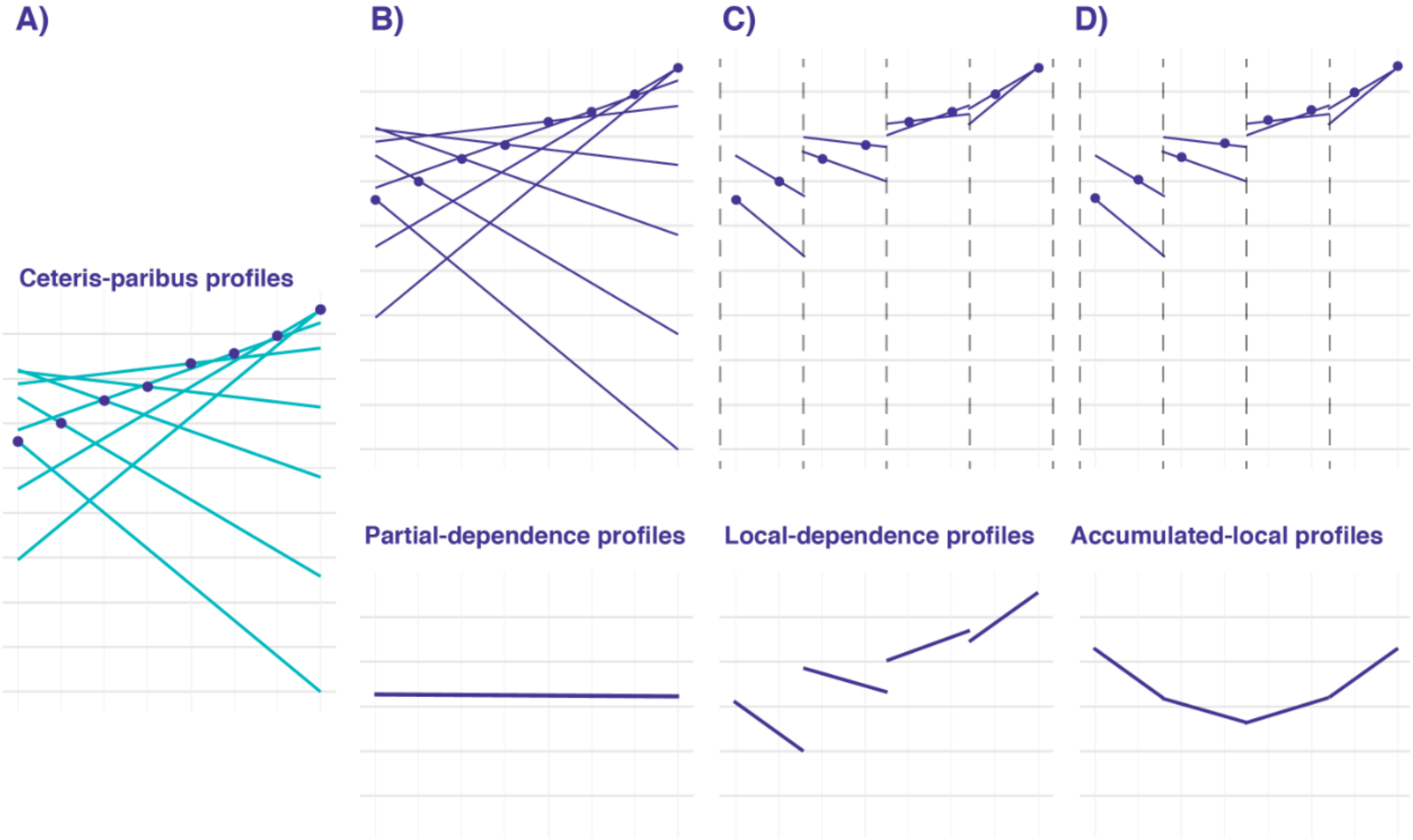

Figure 18.4: Partial-dependence (PD), local-dependence (LD), and accumulated-local (AL) profiles for model (18.10). Panel A: ceteris-paribus (CP) profiles for eight observations from Table 18.4. Panel B: entire CP profiles (top) contribute to calculation of the corresponding PD profile (bottom). Panel C: only parts of the CP profiles (top), close to observations of interest, contribute to the calculation of the corresponding LD profile (bottom). Panel D: only parts of the CP profiles (top) contribute to the calculation of the corresponding AL profile (bottom).

Pros and Cons

When the model is additive, but an explanatory variable is correlated with some other variables, neither PD nor LD profiles will properly capture the effect of the explanatory variable on the model's predictions. However, the AL profile will provide a correct summary of the effect.

When there are interactions in the model, none of the profiles will provide a correct assessment of the effect of any explanatory variable involved in the interaction(s). This is because the profiles for the variable also include the effect of other variables. Comparison of PD, LD, and AL profiles may help in identifying whether there are any interactions in the model and/or whether explanatory variables are correlated.

In most situations, it would be preferred ALE plots over PDPs, **because features are usually correlated to some extent (Molnar et al., 2018).**

Pros and Cons

ALE plots are unbiased means that they work when features are correlated. PDP fail in this scenario because they marginalize over unlikely or even physically impossible combinations of feature values.

ALE plots are faster to compute than PDPs, since the largest possible number of intervals is the number of instances with one interval per instance.

The interpretation of ALE plots is clear

An interpretation of the effect across intervals is not permissible if the features are strongly correlated.

ALE effects may differ from the coefficients specified in a linear regression model when features interact and are correlated.

ALE plots may have small ups and downs with a high number of intervals.

References

Biecek, Przemyslaw. 2018. “DALEX: Explainers for complex predictive models in R.” *Journal of Machine Learning Research* 19 (84): 1–5. http://jmlr.org/papers/v19/18-416.html.

Apley, Daniel W., and Jingyu Zhu. 2020. “Visualizing the effects of predictor variables in black box supervised learning models.” *Journal of the Royal Statistical Society Series B* 82 (4): 1059–86.

Molnar, Christoph, Bernd Bischl, and Giuseppe Casalicchio. 2018. “iml: An R package for Interpretable Machine Learning.” *Journal of Open Source Software* 3 (26): 786. https://doi.org/10.21105/joss.00786.

Please feel free to send e-mail about your questions!

📩 mustafa.cavus@pw.edu.pl